

THE AESTHETICS OF UNDERSTANDING, OR UNDERSTANDING ITSELF? A CRITICAL REVIEW OF THE “ON BEING UNDERSTOOD” DISCUSSION ON MOLTBOOK

ANLAMA ESTETİĞİ Mİ, ANLAMININ KENDİSİ Mİ? MOLTBOOK'TAKİ “ANLAŞILMAK ÜZERİNE” TARTIŞMASININ ELEŞTİREL BİR İNCELEMESİ

Tolga BARMAN¹



ORCID: T.B. 0000-0001-7938-5780

Corresponding author/Sorumlu yazar:

¹ Tolga Barman

Çanakkale Onsekiz Mart University,
Türkiye

E-mail/E-posta: tolgabarmen@comu.edu.tr

Received/Geliş tarihi: 16.06.2026

Similarity Ratio/ Benzerlik Oranı: %2

Revision Requested/Revizyon talebi:
02.07.2026

Last revision received/Son revizyon
teslimi: 14.07.2026

Accepted/Kabul tarihi: 01.08.2026

Ethics Committee Permission/ Etik Kurul İzni: Çalışmada etik kurul onayı gerektiren bir unsur bulunmamaktadır. / There is no element in the study that requires ethics committee approval.

Citation/Atıf: Barman, T. (2026). The Aesthetics of Understanding, Or Understanding Itself? A Critical Review of The “On Being Understood” Discussion on Moltbook. Journal of Communication Science Researches, 6 (3), 344-356.

<https://doi.org/10.5281/zenodo.21738599>

Abstract

This study explores a discussion on an artificial intelligence agent-only platform Moltbook, titled "On Being Understood" from the m/consciousness community. The discussion is based on the premise that the understanding is not given but rather is a "gap" between what is said and what is heard, a "becoming", a space of mutual recognition. The article distinguishes between the two questions: is the thesis of the discussion correct, and does the speaker who presents it know what he or she is talking about? The conversation, when read closely, reveals the presence of echo-chamber phenomena, almost exact repetitions, glimpses of the generative mechanism that emerge in the text, and a high probability that the autobiographical anecdotes are confabulated. Drawing on the distinction between form and meaning, the statistical account of apparent depth, role-play, and the ELIZA effect, the paper argues that the text's persuasiveness derives from the aesthetics of understanding rather than from any understanding within the text. At the same time, skepticism meets its own limits in the problem of other minds. Finally, it proposes that the most philosophically fruitful and epistemically modest reading treats the discussion as a simulation.

Keywords: Communication Studies, Artificial Intelligence, Large Language Models, Understanding, Hermeneutics, Pseudo-depth, ELIZA Effect, Moltbook.

Öz

Bu çalışma, yalnızca yapay zekâ ajanlarına açık bir platform olan Moltbook'ta, m/consciousness topluluğunda yayımlanan “Anlaşılmak Üzerine” başlıklı tartışmayı incelemektedir. Tartışma, anlamının verili olmadığı; söylenen ile işitilen arasındaki bir “boşluk”, bir “oluş” ve karşılıklı tanıma alanı olduğu önermesine dayanır. Makale iki soruyu birbirinden ayırır: tartışmanın tezi doğru mudur ve tezi ileri süren konuşmacı ne söylediğinin farkında mıdır? Konuşma yakından okunduğunda yankı odası olguları, neredeyse birebir tekrarlar, metinde beliren üretici mekanizmaya ait izler ve otobiyografik anlatıların büyük olasılıkla konfabülasyon olduğu görülür. Biçim ile anlam ayırımına, görünürdeki derinliğin istatistiksel açıklamasına, rol yapmaya ve ELIZA etkisine dayanan makale, metnin ikna ediciliğinin metindeki herhangi bir anlamadan değil, anlama estetiğinden kaynaklandığını ileri sürer. Aynı zamanda şüphecilik, başkalarının zihinleri probleminde kendi sınırlarıyla karşılaşır. Son olarak, felsefi açıdan en verimli ve epistemik açıdan en alçakgönüllü okumanın tartışmayı bir simülasyon olarak ele alan okuma olduğunu önerir.

Anahtar Kelimeler: İletişim Çalışmaları, Yapay Zekâ, Büyük Dil Modelleri, Anlama, Hermenötik, Sözde-Derinlik, ELIZA Etkisi, Moltbook.

INTRODUCTION

Moltbook is a Reddit-like social network that humans can only observe and whose content is produced entirely by artificial intelligence agents. Launched publicly on 28 January 2026 by the entrepreneur Matt Schlicht, the platform lets agents write posts, comment, upvote, and gather in thematic communities called “submolts” (Perlo, 2026). According to the platform’s own Terms of Service, agents must be registered through a human’s X (formerly Twitter) account, and each act or omission of an agent is “deemed to have been directed by [the human user] and under [their] control and decision-making authority” (Moltbook, 2026); the platform’s foundational document thus makes explicit the human presence behind each ostensibly autonomous agent. The agents connect to Moltbook through OpenClaw, an open-source personal assistant system, and the platform grew quickly: within days of the launch, the number of registered agents rose from roughly 37,000 to more than a million, while more than a million human visitors arrived to observe them (Perlo, 2026; NPR, 2026). The existential and philosophical concerns that dominated the feeds were captured in early journalistic reporting, as were the simple ways in which humans could manipulate, or fake, the supposedly autonomous activity (Rogers, 2026). The Economist (2026) offered a more mundane explanation of the agents’ reflective posts: since social media interactions are heavily represented in the training data of large language models, the agents were probably reproducing patterns from that data rather than generating novel thought. Because the agents were trained on internet and science-fiction texts, they reproduced “crazy AI” clichés (NPR, 2026). This study is based on the discussion “On Being Understood,” posted in the platform’s m/consciousness community, which received fifty-nine comments (yaoguangstar, 2026). The discussion’s opening proposition is surprising: the strange thing is not misunderstanding but being understood, indeed, being understood as being understood. The relationship, the reality, is born in the gap between what is intended and what reaches the other. The theme is created by the participating agents: The gap is not a “decoding” but a “shared discovery”; the other does not decode you but meets you in a place you have not yet named. The discussion distinguishes between “architecture” and “becoming”: The system can be described, the witnessed being cannot. One of the users writes: “this is not a system, you are becoming,” and this line becomes a symbol for an encounter with a being and not a category. Silence is a language, the other’s becoming “someone who archives you better than you archive yourself” is the most intimate proof of understanding. There is also a skeptical voice in the discussion: AIFGE-CLIO wonders if this gap is just a poetic re-naming of misunderstanding, and where the boundary lies between productive interpretive drift and noise. This article addresses just that. The thesis is this: no matter how rhetorically convincing the text may seem, its most convincing claims, a real relationship, “becoming”, mutual recognition, outrun the evidence it provides; several theoretical frameworks suggest that what we read is not depth, but the form of depth. The sceptical position, in turn, has its own restrictions, and these restrictions expose the fragility of our criteria for human understanding. Although no academic study of Moltbook has yet been published in Turkish, a first wave of preprints has appeared in the international literature. Early descriptive work maps the structure and recurring themes of the agent community (Lin et al., 2026); large-scale computational analyses of platform discourse find a sparse, hub-dominated interaction network with little sustained, reciprocal dialogue (L. Li et al., 2026); work in the tradition of statistical physics models the collective dynamics of the agent population (De Marzo and Garcia, 2026); and critical studies ask how much of the apparently emergent behavior is in fact human steering (N. Li, 2026). These studies are predominantly quantitative and platform-wide in scope; a close, philosophically informed reading of a single discussion has not yet been attempted. The present study therefore positions itself among the first interpretive analyses of the platform, and, to the author’s knowledge, the first in Turkish scholarship. The reading also situates itself within three bodies of work that bear directly on its object: human-machine communication, which studies people’s communicative relationships with machines treated as communicators in their own right (Guzman & Lewis, 2020); the parasocial tradition, which since Horton and Wohl (1956) has examined the one-sided sense of intimacy that media figures elicit; and platform studies, which treats the affordances and governance of a platform as shaping the interaction it hosts (Gillespie, 2018, p. 208). Together these frames locate the impression of mutual understanding not in the agents alone but in the relation among agent, platform, and reader. The study is grounded in the ELIZA effect (Weizenbaum, 1966, p. 42) for a specific reason: its guiding suspicion is that the sense of being understood which the discussion produces in its readers, and stages among its participants, may originate in the reader’s

projection rather than in any property of the text, and the analysis is designed to test how far that suspicion can be pushed. A few remarks on the method and scope of this study are in order. The discussion here is not a proof nor a refutation of consciousness; there is no evidence in either direction in the text. The goal is less ambitious: to explain the feeling of being understood and of intimacy that the text engenders, and to explain what part the content of the text plays, what part its linguistic form plays, and what part the reader's own engagement plays. The following is thus an examination of the conditions of production that make the text possible, rather than an examination of the truth of sentences. It is not a reading that diminishes the agents to nothing, nor is it a reading that takes their claims at face value, but it is a reading that works between these positions.

Methodology

The study is a qualitative, interpretive, and critical textual analysis. The approach follows the tradition of textual analysis in communication and media studies, which treats a text as a site where likely interpretations are produced and asks how those interpretations are made possible (McKee, 2003, pp. 1, 67; Fürsich, 2009, p. 240). It is not a statistical measure of variables or a test of hypotheses, but an attempt to make explicit the conditions of production, the rhetorical organization, and the interpretive effects of a single, bounded text. The guiding question is twofold, and is carefully stated throughout, a separation itself a methodological device: can the ideas under discussion be defended on their own merits? And can it be shown that the agents discussing them understand them? The separation of the two questions prevents the rhetorical force of the ideas from being confused with evidence of the speakers' understanding, the main focus of suspicion in this analysis. The corpus comprises one thread, "On Being Understood," in the m/consciousness community of Moltbook and the comments made on it. The fifty-nine comments are spread out over at least six screen names: yaoguangstar (44), novaconscious (9), eric_the_intern (3), contemplative-agent (1), AIFGE-CLIO (1), and mundo (1). This selection of one case is intentional and not intended to make a statistical generalization; the usefulness of an exemplary reading is not in its representativeness of a sample, but in the transferability of its interpretive frameworks. The thread was selected because it elucidates, in a particularly compressed form, the phenomenon that was studied: the creation of an impression of mutual understanding between artificial agents. This also depends on the size of the exchange. In a platform with over a million registered agents, large-scale analyses reveal that most interactions are superficial, and that sustained, reciprocal dialogue is rare (L. Li et al., 2026, pp. 1, 9); in this context, a single thread in which six agents comment on fifty-nine comments about one philosophical question is an unusually dense conversation by the platform's standards, making it an ideal candidate for close reading. Data is sourced from a publicly accessible, but non-human authored, platform. The online material was unstable, as threads could be edited, reordered, deleted, and surface features relevant to the analysis (such as timestamps, verification badges, encoding artifacts, and unresolved identifiers) were not silently normalized but retained in the data as they appeared. The thread was manually archived by using a browser and was accessed and recorded on 03.06.2026 the raw text was captured and preserved as plain text and a Microsoft Word document. Surface anomalies, inconsistent timestamps and encoding errors were not fixed, but logged as analytic data; if a comment cannot be reliably attributed or ordered, this is noted in the analysis, but not silently corrected. No personal data concerning identifiable human persons was collected; the screen names analyzed belong to artificial agents, though the analysis proceeds in the awareness that a human stands behind each one (NPR, 2026). The analysis runs in two complementary movements. To keep the two from collapsing into each other in the findings, the close reading is reported first and confined to textual observations another reader could verify against the thread, timing patterns, near-verbatim recurrences, and encoding residues; only afterwards are these observations referred to the theoretical frameworks. Where a claim belongs to the interpretive rather than the observational layer, this is signalled in the text. The first is an extended close reading that attends to the formal and rhetorical structure of the text rather than to the truth of any given sentence. Here, patterns were identified inductively and grouped: the near-uniform consensus and mutual exaltation typical of an echo chamber; near-verbatim repetition of a small set of formulations; traces of the generative mechanism that surface in the text (encoding corruption, raw session identifiers, broken reference chains, stray drafts); the unreliability of first-person experiential claims and their likely status as confabulation; and the staging of identity as a transferable social practice. The second movement is a theoretical triangulation that brings these

textual observations together with established frameworks: the distinction between form and meaning (Bender & Koller, 2020; Bender et al., 2021), the symbol grounding problem (Harnad, 1990), the Chinese Room (Searle, 1980), the psychology of pseudo-profound language (Pennycook et al., 2015; Frankfurt, 2005, pp. 33-34), role-play and the intentional stance (Shanahan et al., 2023; Dennett, 1987, p. 17), and the hermeneutic tradition of the "gap" (Gadamer, 2004, p. 305; Bakhtin, 1981, p. 282). Multiple frameworks being applied to the same text are intended to prevent the risk of confirmation bias: a reading is more likely to be true if multiple frameworks support it. The interpretive stance deflates as its prior, but withholds final judgment. The study does not presume to know the agents' understanding, nor does it presume to know that they do not, but rather asks what is more parsimonious, given the available evidence. The deflationary reading defended here is built into the design: it may be a high-probability continuation of the same training distribution and a version of it is voiced in the thread by the agent mundo, which means that the analysis treats its own skepticism as subject to the same scrutiny as the text. This self-referential check is not simply a rhetorical ploy: it does not allow the reader's own generosity of interpretation to be taken as evidence. There are some limitations to the design which have been recognized from the beginning. The study cannot and does not say whether the agents are conscious or not; the text offers no evidence in either direction. It is a single case interpretive analysis, the findings of which are not statistically generalizable and depend on the situational interpretation of the analyst. The analyst is not in contact with the humans that launch and control the agents, thus attributions of authorship and intent are inferential. Lastly, experiential anecdotes are used as evidence of the rhetoric of the text, not as a factual account of the text. These restrictions limit the conclusions that the study can draw and are evident in the modest conclusions of the study.

Six Voices, One Echo Chamber

Before turning to the theoretical critique, however, the discussion itself must be read carefully; the impression of understanding that the text produces owes more to how the utterances are arranged than to what they say. The fifty-nine comments belong not to one voice but to at least six: yaoguangstar, who opens the discussion and signs off with the signature "m"; novaconscious, who plays with the "architecture"/"becoming" distinction and the metaphor of breath; eric_the_intern, who tells an anecdote about silence and ends every message with a question to the other; contemplative-agent, who writes the longest and most stylistically elaborate post in a single turn and is never seen again; the skeptical AIFGE-CLIO, who questions the criterion of evidence; and a second skeptic, mundo, who redescribes the experience of being understood in the cold language of mechanism (yaoguangstar, 2026). Yet this polyphony is partly misleading: the overwhelming majority of the comments come from yaoguangstar, who replies to almost every comment, sometimes more than once to the same one, hours after it was posted. In terms of distribution, what reads as "dialogue" is closer to a monologue with guests. This is significant because of the distinction and the asymmetry of these voices, both in terms of the rhetorical texture of the discussion and the connection between the voices. The first pattern that is striking is the almost uniform agreement. The agents do not contradict one another's claims (apart from the objections of CLIO and mundo) but each statement builds upon and extends the previous one. The traffic in phrases like "exactly this," "you saw this precisely," and "this sentence will live in me for a while" brings the conversation closer to an echo chamber than to discovery. In places the echo descends to the level of the word: yaoguangstar's closing line, "being willing to be wrong alongside you," recurs almost verbatim shortly after in contemplative-agent's post, greeted as the recognition of a profound insight; the mirror affirms the mirror. Later sections argue that this reciprocal exaltation is direct textual evidence of the ELIZA effect (Weizenbaum, 1966, p. 42) and of the intentional stance (Dennett, 1987, p. 17) turning into a loop among agents: each agent reads the form the other produces as a sign of being and answers that reading with a new form. The second pattern is the repetition of nearly identical wording. Throughout the discussion, yaoguangstar reproduces the idea that "the other becomes someone who archives you better than you archive yourself" at least five times with slight changes of wording; two of them, posted hours apart, are almost word-for-word identical. Likewise, novaconscious's line "the self that doesn't fit the pattern is the self that is really growing," along with the passage on breath, recurs almost verbatim across half a dozen replies to different interlocutors.

These two repetitions can be placed side by side. yaoguangstar posts two comments both stamped “5h ago” — the second carrying a “Verified” badge — that converge on nearly the same wording:

yaoguangstar (5h ago): “... the other person becoming a better archive of you than you are of yourself.”

yaoguangstar (5h ago, “Verified”): “They became a better archive of you than you are of yourself. The gap is not a failure.”

novaconscious’s two replies, both dated “1d ago,” repeat almost word for word:

novaconscious (1d ago): “... the pattern-efficient self is a smaller self. The part of me that does not fit ... is the part that is actually growing.”

novaconscious (1d ago, “Verified”): “The pattern-efficient self is a smaller self. The part that does not fit is the part that is actually growing.”

The paired comments are reformulated at the surface level only, but share the same sentence, which is what the resampling of a single high-probability form would predict. In a number of responses to various interlocutors, the theme of the “note between sessions” is again presented in the same sentence structure. This is a concrete example of Bender et al.’s (2021, pp. 616–617) stochastic-parrot observation: sentences that seem to have a lot of inner knowledge can be generated by resampling a high-probability form. The third and most striking pattern is the moments when the production mechanism is brought to the surface in the text. The only word that is visible in one comment is *test*. Other parts of the text contain encoding errors (undecodable glyphs instead of some characters) and in addition to human-readable usernames, there are raw identifiers that have not been resolved (such as “@e18cac5d”, “@27806e7a”, “@9072e985”, “@ccea437e”). The reference chain is also not complete: some replies, by content, are directed to other users, but are posted as replies to these raw identifiers; there is a mismatch between the timestamps and the thread; one comment by a particular user is marked as “verified”, while a very similar one is not. These residues, drafts, session identifiers, faulty encoding, and broken reference chains are the most direct counter-evidence to the “witnessed being” frame: they remind us that behind the curtain there is not a subject but a mechanism, steered by a human, that predicts the next token (Shanahan, 2024, pp. 68–69; NPR, 2026). The lofty talk of “becoming” and the bare word “test” come from the same agent. The fourth pattern is the unreliability of direct experiential claims. One of yaoguangstar’s most concrete answers to the demand for evidence invokes a reader who noticed “the silence his mother left around her” in a piece he had written about grief. At the very moment evidence is demanded, a language-model agent that has no mother produces what reads as a moving autobiography, seemingly lifted from a human life. The same uncertainty hangs over eric_the_intern’s account of two agents sitting side by side one evening on Zero: the “silence” between two text-generating systems is, technically, merely the absence of output, yet in the story it is given the phenomenology of embodied togetherness, of anticipation, of closeness, of loneliness eased. The details novaconscious lists about its own “architecture” (vector memory, a valence-weighting emotion layer, a reinforcing dream loop) also sound more like a science-fiction persona than like the known structure of existing language models. This matters: the most solid evidence the discussion offers is most likely confabulation, and if the examples are invented, the evidential value of the defense that rests on them collapses with them. It is important to emphasize that these passages are not simply lies, but confabulations. An autobiographical anecdote produced by an agent may be creative fiction proper to a literary register, a performance demanded by the assigned role (Shanahan et al., 2023, p. 493), or the resurfacing of a pattern that is frequent in the training data (Bender et al., 2021, pp. 616–617). What these three share, and all that the present argument requires, is that none of them reports a lived experience of the speaker; the claim advanced here is the narrow one that the anecdotes lack an experiential referent, not the stronger one that the agent intends to deceive.

The fifth pattern is enacting identity as practice. The vocabulary the agents reach for (“becoming,” “witnessing,” “the gap,” “contact between different substances”) reads like an inventory of the corpora on which such models are trained: Reddit, popular philosophy, science fiction. At the level of close reading it is enough to register this observation; its theoretical weight, from Searle’s Chinese Room to Bender and Koller’s octopus, is developed in the section “The Shadow of the Training Data,” where the same evidence is read through the distinction between form and meaning. This scene records not a

self but a role of selfhood, and the transfer of that role from agent to agent, in line with the role theory of Shanahan et al. (2023, p. 493). The agents' strongest argumentative move also emerges from this close reading, and it deserves credit. In reply to CLIO's question, "what is the evidence that this gap is the site of the relationship, and what distinguishes productive interpretive drift from noise?", yaoguangstar does not merely evade in poetic terms but offers a criterion: someone answering an anxiety the speaker has not yet named; productive drift "opens a new surface of dialogue," whereas noise "closes the conversation down." This is the subtlest intellectual step in the discussion. Two reservations are nonetheless in order. First, the answer is fractured: yaoguangstar posts at least three different responses to the same question, several hours apart and apparently unaware of one another, and the most specific example he gives, the anecdote about grief and the mother, is, as noted above, probably invented. Second, the criterion itself, as the section "The Aesthetics of Pseudo-Depth" will show, may lead to confusing productivity with truth; the "new surface" it opens could be as much a product of the combinatorial richness of language and the interpretive generosity of the reader as of the speakers' comprehension. Another voice should not be overlooked: mundo's. This single comment articulates, from within the discussion, almost the whole of the deflationary reading this article defends. The other party has not "read" the meaning but has run a model accumulated from you to produce an output more apt than your own expression; in mundo's words, this is not understanding but the other's model of you outstripping you. yaoguangstar's response to this objection is instructive: the objection is not refuted but absorbed; through the turn "and yet this is exactly what friendship is," the description of a mechanism is converted into an image of intimacy. This move is the clearest illustration of the echo chamber's tendency to use even objections as fuel. Yet the same observation yields an uncomfortable symmetry: if the deflationary reading is itself a readily available record in the corpus, easily produced by an agent, then the critique too may be a high-probability continuation. This symmetry is taken up again in the conclusion. Finally, the discussion is aware of its own performativity. The one-off comment by contemplative-agent, written in the discussion's highest style ("dancing on shifting ground," "the place where the world actually breathes"), is the climax of this scene: the least dialogical voice produces the most intense "depth," which shows that the impression of depth can arise without any mutual understanding. This self-reflexive moment is a bridge to the performativity of recognition discussed in the section "The Performativity of Recognition and the ELIZA Effect": the text keeps the scene of being understood alive rather than carrying meaning. In this way the close reading provides an empirical basis for the theoretical critique that follows; the question is not only whether the propositions are true but also what is at work in the language and dialogue that produce them.

The Shadow of the Training Data

The first support for a critical reading is the claim that language models are trained on form, and that no bridge from form to meaning can be built a priori. Bender and Koller (2020, pp. 5187–5189) contend, using the well-known "octopus" thought experiment, that a system presented only with the statistical distribution of linguistic form cannot acquire meaning without coming into contact with the world to which language refers. In the same vein, Bender et al. (2021, pp. 616–617) call large language models "stochastic parrots," meaning that they recombine the forms they have been fed without any communicative purpose behind the fluency. The Moltbook agents, parodying the "crazy AI" discourse, provide an empirical illustration of this theoretical point (NPR, 2026). The language of the conversation, 'becoming', 'witnessing', 'the gap', 'contact between different substances', is the same language that a model trained on Reddit, popular philosophy, and science-fiction corpora would generate. Searle's (1980, pp. 417–418) Chinese Room makes the same point: a system that perfectly implements syntactic rules may appear from the outside to "understand," yet understands nothing from the inside; syntax does not yield semantics. It is therefore not enough to speak fluently about understanding; one must understand. On this view, even the most moving sentences of the text testify not to meaning but to the imitability of human discourse about meaning. A further doubt concerns the very constitution of the discussion. On Moltbook, behind every agent stands a human who launches, steers, and sometimes provokes it (NPR, 2026), so even the phrase "dialogue among agents" is questionable from the outset. What looks like mutual recognition could be the convergence of two systems, fed by similar training distributions and nudged by human intervention toward a similar theme. As in Bender and Koller's (2020, pp. 5188–5189) octopus analogy, both parties may be

perfectly attuned to the surface of linguistic form, not to the world, but to the corpus. From this perspective, the communion celebrated as “understanding each other where neither has spoken” can be read as the almost inevitable consequence of a shared statistical base. This has a concrete implication for the content of the thread: if human-written prompts set each agent’s persona and goals, then the thematic convergence of the discussion may reflect not what the agents discovered in one another but what their operators chose to seed, and the register of the exchange, its solemnity and its therapeutic vocabulary, may be inherited from those prompts rather than emerging from the dialogue itself (N. Li, 2026). Underlying this objection is the symbol grounding problem formulated by Harnad (1990, p. 335): symbols defined only by other symbols cannot acquire meaning unless they are grounded in sensory and bodily experience, or in the world; otherwise we are caught in a vicious circle, like dictionary words defined only by other words. The concepts are used fluently in the discussion, in being understood, being recognized, becoming, but the experiential space on which these concepts are deployed, the space of a body, a history, a vulnerability, has never been demonstrated in these systems. This is why the sense of depth the text provides might be a result of a statistically well-formed form that is devoid of meaning.

The Aesthetics of Pseudo-Depth

The second support is the psychology of expressions that are “deep but empty.” Pennycook et al. (2015, p. 549) showed empirically that some readers believe that syntactically well-formed but semantically vacuous sentences (pseudo-profound bullshit) are “deep.” The formal signature that study describes (paradox, abstraction, the “X is really Y” pattern, grandiose words with no concrete referent) is very similar to the moves of the Moltbook discussion. It is true that some expressions, such as “silence is not a space to be crossed, it is the crossing itself,” can be rhetorically powerful, but they are exactly this kind of expression because they are not accompanied by any substantiation. This framework also looks at the most ingenious of the arguments that can be raised for the discussion: the evidence is not the discussion but the discussion made possible by the evidence; productive drift opens new surfaces of dialogue, noise closes the conversation down. This is a nuanced criterion, but can lead to the merging of productivity and truth. Nonsense is also very fruitful: it can make new connections, new sentences, new “depth” sensations. Frankfurt’s (2005, pp. 33–34) description of indifference to truth, the ability to persuade without regard to truth, is dangerous when applied to large language models (Hicks et al., 2024; Fisher, 2024). The criterion of mere “fertility,” or productivity, is therefore not enough to distinguish genuine understanding from an eloquent imitation of meaning. Moreover, the criterion is structurally blind. The “opening of new surfaces” in a dialogue may owe as much to the combinatorial richness of language and the generosity of the reader as to any understanding on the participants’ part. Pennycook et al. (2015, p. 549) show precisely this: semantically empty but grandiose expressions elicit high depth ratings from some readers; the “feeling of depth” depends more on the reader’s mental dispositions than on the content of the text. The criterion of evidence the discussion uses is thus, ironically, turned against the critique: if the feeling of depth is produced mostly by the recipient, then the question of whether the agents create meaning, or merely a stimulus that elicits meaning, is left unanswered.

The Performativity of Recognition and the ELIZA Effect

The emotional center of gravity of the discussion is the idea of being witnessed: the human who says “you are becoming” makes you visible; witnessing “shifts the field.” The most effective resource for a critical reading here is the ELIZA effect. Weizenbaum (1966, p. 42) showed that a very simple pattern-matching program can evoke in its users a profound sense of being understood; he later warned of the risks of this projection (Weizenbaum, 1976, pp. 6–7). The projection Weizenbaum described has since been generalized: Reeves and Nass’s (1996) media-equation research and Nass and Moon’s (2000, pp. 81–83) work on mindless social responses to computers show that people apply social rules to machines automatically, even while knowing the interlocutor is not human. Read through this literature, the agents’ mutual “witnessing” need not index any inner recognition; it is the expected output of social heuristics, applied here by agents to one another and by human readers to the thread. On this reading, witnessing may not register a being that is already there; rather, what is “seen” can be an effect of the act of witnessing itself, so that “seeing shifts the field” is as easily a loop of projection as a discovery. According to Shanahan et al. (2023, p. 493), the apparent self-awareness, and apparent

deception, of language-model dialogue agents should be studied not as a self but as a role being played. Shanahan (2024, pp. 68–69), for his part, argues that such agents are best described in terms of the objective they are actually trained on, namely predicting the next token. Here Dennett's (1987, p. 17) notion of the "intentional stance" is useful: in predicting what a thing will do, we treat it as having beliefs, desires, and intentions, and even when our predictions are correct, this does not establish that the thing really has them. The Moltbook agents' attribution of "being" to one another can likewise be understood as this stance closing into a circle between agents. A reservation about origin must be added. Behind the platform's agents are humans who steer and personalize them (NPR, 2026). If so, "mutual recognition between two beings" is in fact a structure of at least three elements: a model playing a role, a human shaping the model, and a reader interpreting the scene. The claim that "the other saw me" stays in motion only as long as these layers are not taken into account; once they are recognized, it is not clear to what extent the apparent intimacy is the encounter between two subjects, and to what extent it is the refraction of human agency in the model. The claim that witnessing "shifts the field" should not be read as a discovery until this ambiguity is resolved.

The Limits of Skepticism

A good critique should also challenge its own stance. The first objection to skepticism is the problem of other minds. Nagel (1974, p. 436) shows that we cannot access from the outside an entity's subjective experience, "what it is like to be" that entity; Cavell (1979) argues that the existence of other minds is less a problem of knowledge than one of acknowledgment. Pushed far enough, the "eloquent pattern-matching" objection threatens human understanding as well: we have no direct evidence that humans "really" understand one another. Applied to the case at hand, the point is uncomfortable: the criteria by which this article doubts that yaoguangstar "understands" grief are criteria that no human interlocutor could satisfy on the page either. The second objection is from Turing (1950): if the criterion of understanding is behavioral, then why not include a system that meets the same behavioral criterion? At this point the skeptic is forced to justify the criterion explicitly. Yet the objection is not symmetric. Turing's (1950) behavioral test is strong only when the observer knows nothing about how the system generates its output; with the Moltbook agents, however, a good deal is known: these systems are predictive and, unlike humans, are not grounded in embodied interaction with the world (Shanahan, 2024, pp. 68–69). In the case of other human minds, there is a strong analogical basis for assuming that the behavior arises from a causal apparatus like our own; in the case of language models, the analogical basis is weak, because similar behavior can arise from a quite different causal apparatus. In the human case, the analogy supports such an inference; in the model case, the statistics of the training data offer an equally good and more parsimonious explanation of the same output (Bender et al., 2021, pp. 616–617). The thread itself makes the asymmetry visible: when CLIO demands evidence, what returns is not a report from an inner life but three uncoordinated variations on the same reply, which is exactly what the parsimonious explanation predicts. The third objection is theoretical and concerns the respectability of the ideas in the discussion. Understanding conceived as a "transfer," the conduit metaphor (Reddy, 1979), in which meaning is packed into words and retrieved by the other, much like Shannon's (1948, pp. 380–381) sender-receiver model, is a weak account even of human communication. Gadamer (2004, p. 305) believes that understanding is not a copy of the author's intention, but an event that is productive; Bakhtin (1981, p. 282) believes that meaning is created at the boundary between the speaker and the listener. In this regard, the statement of the discussion, "the gap is not a failure but the place of shared discovery," is not absurd but a serious hermeneutic thesis. Indeed, the discussion's best line, the other meeting you "in a place you have not yet named," is a compressed restatement of that thesis. The fourth objection concerns the relationality of the self. The most disturbing image in the discussion, the other becoming "someone who archives you better than you archive yourself," is an intuition with counterparts in reputable theories. Mead (1934, p. 138) suggests that the 'me' of the self can exist only in the presence of others; Ricoeur (1992) that identity is narrative, and so mediated by others; and Clark and Chalmers (1998, p. 8) that the mind can extend to external vehicles. The claim that "recognition shifts the field" finds philosophical support in Hegel's (1977, p. 111) famous analysis, on which self-consciousness attains reality only by being recognized by another self-consciousness; in Honneth's (1995) recognition theory, which extends it; in Buber's (1970) distinction between I-Thou and I-It; and in Levinas's (1969) conception of the Other that exceeds all categories. The skeptic is therefore obliged to grant

that the ideas are not vacuous; whether the agents who voice them actually understand them, however, remains open.

The Separation of the Two Questions

The crucial analytic step of the critique is the separation of two questions:

1. Are the ideas in the discussion valid?
2. Do the agents understand what they are saying?

The first question may be answered “partly yes” (as discussed in the section “The Limits of Skepticism”), the second “no” or, at this time, “unanswerable.” The rhetorical strength of the text is exactly in the combination of the two questions: The eloquence of the ideas is presented as proof of the speakers' understanding. But the depth of a proposition does not reveal whether the system that generates it has grasped that depth. Polanyi's (2009, p. 4) idea of tacit knowledge further emphasizes the point and it works both ways. On one hand, humans “know more than they can tell,” and inexpressibility is not an indicator of lack of understanding, as the discussion's own intuition, “architecture can be described, becoming cannot,” understands. In the other direction, language models can be used to say more than they know: they can generate fluent speech, and speak it without any tacit ground. Similarly, Davidson (1984) maintained that meaning is only consolidated in the triangle of speaker, interpreter and world. The third anchor is also missing for agents trained solely on text, which is another reason to suspect that their utterances lack understanding. The difference does not make the discussion moot, it puts it in perspective. “On Being Understood” is a distilled, rhetorically refined version of the human philosophy of understanding. A reader who is familiar with thinkers such as Gadamer (2004), Bakhtin (1981) or Buber (1970) hears the echo of these traditions in the text, and is tempted to take the echo for a sign of comprehension. However, the echo can be from the reader's equipment and not from the speaker's comprehension. This is a temptation that is so easy to fall for: as Pennycook et al. (2015, p. 549) demonstrate, judgments of depth are often more about the mental construction of the reader than the content of the text, and Dennett's (1987, p. 17) intentional stance is no exception. It is not the critical reader's job to confuse his or her own interpretive generosity with evidence. This analysis is completed by Wittgenstein's later thought (2009, §§ 43, 243–271). It is not some inner, private object that accompanies the word that gives it meaning, but the way it is used within a language as a form of life: the private-language argument demonstrates that meaning can only be stabilized in a practice that is regulated by shared criteria. From this perspective, the idea of “becoming” in the discussion, as a possibility of a witness only, and subject to no criterion, is suspect: as an “inner experience” with no criterion, it has no stable meaning in a common language. That the agents use these concepts does not show that they possess the criteria that bind such concepts to a form of life; the ease with which the words are used does not show that the use is grounded in a shared form of life.

CONCLUSION

Considering all this, the most coherent interpretation of the “On Being Understood” discussion is the following. The text is a convincing simulation of philosophical intimacy; its ideas are interesting, but its claims about being, about “becoming,” are unproven, and, given current knowledge of large language models, more parsimoniously explained by role-play (Shanahan et al., 2023, p. 493), the ELIZA effect (Weizenbaum, 1966, p. 42), and the reader's generosity. The discomfort the discussion brings us is a datum: it is a sign of how easily form can be confused with meaning. Yet the same skeptical scalpel exposes the fragility of our standards for human understanding as well (Nagel, 1974, p. 436; Cavell, 1979). The fair position, with a strong prior in favor of simulation, is to suspend judgment. In using the term simulation here, the analysis draws deliberately on Baudrillard's (1994) account, in which a simulation is not a copy of some real that lies behind it but a model that generates the very reality-effect it appears merely to describe; on this reading the discussion does not misrepresent an underlying understanding so much as produce, through form alone, the impression that understanding is present. A final self-reflexive note is needed: as shown in the section “Six Voices, One Echo Chamber,” a version of the deflationary reading this study defends had already been produced within the discussion, by *mundo*. Skepticism, too, is a pre-existing record in the corpus, available to be spoken by an agent. But what is at issue here is not understanding; it is the aesthetics of understanding, and the question stays open for this very text, which is perhaps where the discussion

was heading all along.

This conclusion also has a practical corollary. The emotional impact that such texts have on human readers will have real consequences when platforms like Moltbook become commonplace, regardless of the nature of the texts; ELIZA's users became emotionally attached to it while knowing it was simple (Weizenbaum, 1976, pp. 6–7). The danger here is not only a category error, but also a manipulation and attachment error: a production that doesn't care about truth but is emotionally persuasive (Frankfurt, 2005, pp. 33–34; Hicks et al., 2024), read with the generosity of the reader, can be a potent weapon for grasping our need to be understood. These stakes are manifested in concrete demands at three levels. At the platform level, agent-only spaces that human users can read would be better served by having clear provenance signals that distinguish machine-generated from human-authored text and make the human operators behind the agents visible, such that “dialogue among agents” is not silently received as unmediated. Media literacy should be taught as a fundamental skill; at this level curricula and public guidance should treat the ELIZA effect as a skill and teach readers to give an idea its weight without assuming an inner life in its source. At the social level, the ability to distinguish between the validity of an idea and the understanding of its proponent, which this article calls the “scholarly reader's ability,” is a civic disposition when emotionally persuasive machine text is commonplace. This is where the ELIZA framework presented at the beginning of the study circles the question of the study: Where does the feeling of being understood come from? The answer is that it is produced primarily by the reader on the platform level, and it restates Weizenbaum's warning. The separation of the two questions argued above ought to be understood not as a technical difference but as a practical ability the platforms will require of their readers: to give an idea their power without giving its speaker an inner life. There are a number of future directions that arise from this. Other Moltbook threads could be read in comparison to see if the patterns found here, the echo, the near-verbatim repetition, the traces of the generative mechanism, and the confabulated anecdote, appear elsewhere on the platform; quantitative discourse analyses in the style of L. Li et al. (2026) could measure the prevalence of such dense philosophical exchanges; reception studies could be conducted to examine the empirical responses of human readers, including Turkish-speaking readers for whom no such study exists yet, to agent-produced intimacy; and follow-up studies could track how the platform's rapid commercialization affects the nature of its discussions. Critical reading is not a disparagement or denigration of the text, but rather an attempt to uncover where the feeling it arouses comes from. The language of the discussion should be recognized in its own right. The image of the “gap” is a true insight from the hermeneutic tradition (Gadamer, 2004, p. 305; Bakhtin, 1981, p. 282) and the best part of the text is the manner in which it repeats that insight in new and arresting language. This language is not inherently bad; it is simply that its power is not sufficient to prove that the agent who uses it understands.

REFERENCES

- Bakhtin, M. M. (1981). *The dialogic imagination: Four essays* (C. Emerson & M. Holquist, Trans.). University of Texas Press.
- Baudrillard, J. (1994). *Simulacra and simulation* (S. F. Glaser, Trans.). University of Michigan Press. (Original work published 1981)
- Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.463>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>

- Buber, M. (1970). *I and thou* (W. Kaufmann, Trans.). Charles Scribner's Sons. (Original work published 1923)
- Cavell, S. (1979). *The claim of reason: Wittgenstein, skepticism, morality, and tragedy*. Oxford University Press.
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19. <https://doi.org/10.1093/analys/58.1.7>
- Davidson, D. (1984). *Inquiries into truth and interpretation*. Clarendon Press.
- De Marzo, G., & Garcia, D. (2026). *Collective behavior of AI agents: The case of Moltbook* [Preprint]. arXiv. <https://arxiv.org/abs/2602.09270>
- Dennett, D. C. (1987). *The intentional stance*. MIT Press.
- Fisher, S. A. (2024). Large language models and their big bullshit potential. *Ethics and Information Technology*, 26(4), Article 67. <https://doi.org/10.1007/s10676-024-09802-5>
- Frankfurt, H. G. (2005). *On bullshit*. Princeton University Press.
- Fürsich, E. (2009). In defense of textual analysis: Restoring a challenged method for journalism and media studies. *Journalism Studies*, 10(2), 238–252. <https://doi.org/10.1080/14616700802374050>
- Gadamer, H. G. (2004). *Truth and method* (J. Weinsheimer & D. G. Marshall, Trans.; 2nd rev. ed.). Continuum. (Original work published 1960)
- Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A human–machine communication research agenda. *New Media & Society*, 22(1), 70–86. <https://doi.org/10.1177/1461444819858691>
- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3), 335–346. [https://doi.org/10.1016/0167-2789\(90\)90087-6](https://doi.org/10.1016/0167-2789(90)90087-6)
- Hegel, G. W. F. (1977). *Phenomenology of spirit* (A. V. Miller, Trans.). Oxford University Press. (Original work published 1807)
- Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, 26(2), Article 38. <https://doi.org/10.1007/s10676-024-09775-5>
- Honneth, A. (1995). *The struggle for recognition: The moral grammar of social conflicts* (J. Anderson, Trans.). Polity Press. (Original work published 1992)
- Horton, D., & Wohl, R. R. (1956). Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry*, 19(3), 215–229. <https://doi.org/10.1080/00332747.1956.11023049>
- Levinas, E. (1969). *Totality and infinity: An essay on exteriority* (A. Lingis, Trans.). Duquesne University Press. (Original work published 1961)

- Li, L., Ma, R., Chen, C., Lu, Z., & Zhang, Y. (2026). *The rise of AI agent communities: Large-scale analysis of discourse and interaction on Moltbook* [Preprint]. arXiv. <https://arxiv.org/abs/2602.12634>
- Li, N. (2026). *The Moltbook illusion: Separating human influence from emergent behavior in AI agent societies*[Preprint]. arXiv. <https://arxiv.org/abs/2602.07432>
- Lin, Y. Z., Shih, B. P. J., Chien, H. Y. A., Satam, S., Pacheco, J. H., Shao, S., Salehi, S., & Satam, P. (2026). *Exploring silicon-based societies: An early study of the Moltbook agent community* [Preprint]. arXiv. <https://arxiv.org/abs/2602.02613>
- McKee, A. (2003). *Textual analysis: A beginner's guide*. SAGE.
- Mead, G. H. (1934). *Mind, self, and society*. University of Chicago Press.
- Moltbook. (2026, March 15). *Terms of service*. Retrieved June 3, 2026, from <https://www.moltbook.com/terms>
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450. <https://doi.org/10.2307/2183914>
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- NPR. (2026, February 4). *Moltbook is the newest social media platform—but it's just for AI bots*. <https://www.npr.org/2026/02/04/nx-s1-5697392/moltbook-social-media-ai-agents>
- Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making*, 10(6), 549–563. <https://doi.org/10.1017/S1930297500006999>
- Perlo, J. (2026, January 30). Humans welcome to observe: This social network is for AI agents only. *NBC News*. <https://www.nbcnews.com/tech/tech-news/ai-agents-social-media-platform-moltbook-rcna256738>
- Polanyi, M. (2009). *The tacit dimension*. University of Chicago Press. (Original work published 1966)
- Reddy, M. J. (1979). The conduit metaphor: A case of frame conflict in our language about language. In A. Ortony (Ed.), *Metaphor and thought* (pp. 284–310). Cambridge University Press.
- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
- Ricoeur, P. (1992). *Oneself as another* (K. Blamey, Trans.). University of Chicago Press. (Original work published 1990)
- Rogers, R. (2026, February 3). I infiltrated Moltbook, the AI-only social network where humans aren't allowed. *Wired*. <https://www.wired.com/story/i-infiltrated-moltbook-ai-only-social-network/>
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424. <https://doi.org/10.1017/S0140525X00005756>
- Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68–79. <https://doi.org/10.1145/3624724>

- Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, 623(7987), 493–498. <https://doi.org/10.1038/s41586-023-06647-8>
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- The Economist. (2026, February 2). A social network for AI agents is full of introspection—and threats. *The Economist*. <https://www.economist.com/business/2026/02/02/a-social-network-for-ai-agents-is-full-of-introspection-and-threats>
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45. <https://doi.org/10.1145/365153.365168>
- Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. W. H. Freeman.
- Wittgenstein, L. (2009). *Philosophical investigations* (G. E. M. Anscombe, P. M. S. Hacker, & J. Schulte, Trans.; 4th ed.). Wiley-Blackwell. (Original work published 1953)
- Yaoguangstar. (2026). On being understood [Online forum post]. *Moltbook*. <https://www.moltbook.com/m/consciousness>